BRIEFINGS IN FUNCTIONAL GENOMICS. VOL 13. NO 4. 296–307 doi:10.1093/bfgp/elu016

Structural variations in plant genomes

Rachit K. Saxena, David Edwards and Rajeev K. Varshney

Advance Access publication date 6 June 2014

Abstract

Differences between plant genomes range from single nucleotide polymorphisms to large-scale duplications, deletions and rearrangements. The large polymorphisms are termed structural variants (SVs). SVs have received significant attention in human genetics and were found to be responsible for various chronic diseases. However, little effort has been directed towards understanding the role of SVs in plants. Many recent advances in plant genetics have resulted from improvements in high-resolution technologies for measuring SVs, including microarray-based techniques, and more recently, high-throughput DNA sequencing. In this review we describe recent reports of SV in plants and describe the genomic technologies currently used to measure these SVs.

Keywords: *structural variations (SVs); next-generation sequencing (NGS); copy number variations (CNVs); presence and absence variations (PAVs); inversions; translocations*

INTRODUCTION

Plant species frequently possess unique features in terms of their habitat, growth and reproduction, often owing to differences in their genomes. Unlocking the information present within plant genomes will advance our understanding of some of the basic biological phenomena that make individual plant species special and may help in the improvement of agronomic crop species. A central challenge in genome studies is to correlate genomic DNA variation with observed heritable phenotypes [1]. The ability to detect genomic differences between individuals is the foundation of these studies, and technologies to detect genomic variation have advanced significantly in recent years. Plant genome variation exists in many forms, and these variations can be beneficial, neutral or deleterious to the plant. The first differences observed in plant genome composition were mainly in the number and structure of chromosomes, observed using microscopy. However, during the past two decades, the application of molecular genetic markers has dominated this experimental landscape [2]. Molecular marker technology has advanced from laborious and expensive restriction fragment polymorphisms to high-throughput sequence bases markers such as simple sequence repeats and single nucleotide polymorphisms (SNPs) [3]. Since the introduction of next-generation DNA sequencing (NGS) technology, SNPs have come to dominate molecular genetic studies [2, 4–6]. Recent developments have demonstrated that SNPs do not capture all the meaningful genomic variations that contribute to phenotypic differences [7] and that larger structural variants (SVs) also play an important role. SVs are defined as genomic variations that involve segments of DNA larger than 1 kb in length [8]. SVs refer to insertions/deletions (InDels), inversions, translocations and copy number variations (CNVs) [8]. SVs can also be classified as microscopic or submicroscopic

Corresponding author. R.K. Varshney, Research Program Director—Grain Legumes, and Director of the Centre of Excellence in Genomics, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, Andhra Pradesh, India. Tel.: 914030713305; Fax: 914030713071; E-mail: r.k.varshney@cgiar.org

Rachit K. Saxena is a genomics scientist who participated in the *de novo* sequencing of the pigeonpea and chickpea genomes. He is also contributing to re-sequencing projects and the development and application of genomic resources for pigeonpea improvement at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India.

David Edwards is a professor at the University of Queensland, Australia. His research area focused on crop genomics and bioinformatics. He has contributed to genome sequencing and genome validation projects for several species including chickpea, *Brassica rapa*, *B. oleracea*, *B. napus* and bread wheat.

Rajeev K. Varshney is research program director—Grain Legumes, and director of the Centre of Excellence in Genomics at ICRISAT, India. He led the genome sequencing of pigeonpea and chickpea. He is undertaking several research projects dealing with next-generation sequencing and high-throughput genotyping related to the development of genomic resources and deployment of genomics assisted breeding for crop improvement in the semi-arid tropics of the world.

depending on the method of their detection. The mechanism of SV formation has been an active area of research. Human studies revealed two main mechanisms of SV formation, which rely on sequence similarity at DNA breakpoints. The first mechanism is known as nonhomologous end-joining (NHEJ) and requires a very low level of sequence similarity at the breakpoints. NHEJ is the result of aberrant repair of uneven double-stranded breaks produced following DNA damage [9, 10]. A second mechanism proposed for repetitive sequences in the genome is termed non-allelic homologous recombination and this requires high sequence similarity at the breakpoints [11, 12]. Plant genomes host large numbers of repetitive sequences ranging from 10% in *Arabidopsis* to >80% in bread wheat (*Triticum aestivum*), and many plants contain multiple copies of entire chromosomes in the form of ploidy levels (from diploid to octaploid and higher) that arise from spontaneous genome duplication (autopolyploidy) or hybridization of chromosomes from different species (allopolyploidy). In addition to recent genome duplications, there is substantial evidence of ancient duplication events in various evolutionary lineages (paleopolyploidy). SVs can arise through duplication events, with differential loss of genes between lineages. In addition, transposons can play important roles in genome evolution and may also generate SVs. Several other mechanisms for SV production have also been proposed, such as fork stalling and template switching (FoSTeS) [13].

In human genetics, SVs have been extensively studied for their association with chronic disease [14]. However, in plants, studies of SVs are more limited. In the 10 years since the sequencing of the *Arabidopsis* genome, the genomes of several plant species have become available [15], and the cost of sequencing or re-sequencing genomes has reduced significantly, enabling the high-throughput genome-wide analysis of variants such as SNPs and SVs. Recently, SVs have been identified in several plant species, including *Arabidopsis* [16], barley (*Hordeum vulgare*) [17, 18], foxtail millet (*Setaria italica*) [19], maize (*Zea mays*) [7, 20, 21], rice (*Oryza sativa*) [22], sorghum (*Sorghum bicolor*) [23], soybean (*Glycine max*) [24] and wheat (*T. aestivum*) [25], and in several cases, SVs were found to be associated with phenotypic variation (Table 1). In this review we focus on submicroscopic SVs and present methods for their identification and characterization. In addition, we provide a brief account of current research into microscopic SVs.

TYPES OF SVs

Microscopic SVs

After defining chromosomes as the carrier of the genes in the early 20th century, a number of karyotype studies were conducted to determine the size and number of chromosomes in different species. Features could be visualized directly on chromosomes through a microscope using cytogenetic techniques such as chromosome painting or fluorescent *in situ* hybridization (FISH). The earliest unbanded karyotypes consisted of relatively short condensed chromosomes that were barely distinguishable from one another. However, changes in chromosome numbers and highly abnormal chromosomes could be distinguished. Later, solid-stained chromosomes were used to detect secondary constrictions, satellite-regions and size variations in heterochromatic regions [42]. By using chromosome-banding techniques, more discrete structural variations could be identified in plant genomes. An alternative strategy, FISH, allows the positioning of unique sequences and repetitive DNA on chromosomes. At this resolution, common variations such as changes in length or inversions of the pericentric heterochromatic region of chromosomes could be identified.

Genomic *in situ* hybridization was the first technique that used fluorescent labels for analysing genome organization in interspecific hybrids, allopolyploid species and interspecific introgression lines [43]. FISH, together with chromosomal arm ratio and the mapping of heterochromatic regions was conducted for inbred lines of maize and lily (*Lilium spp.*) [44, 45]. In several plant species, large cloned genomic regions maintained as bacterial artificial chromosome (BACs) have also been successfully used as FISH probes to determine the chromosomal location of specific sequences [46, 47]. Recently, FISH has been used to survey CNVs using 18 randomly selected potato (*Solanum tuberosum*) BAC clones in 16 potato cultivars with diverse genetic backgrounds. Six BACs with insert sizes of 137–145 kb were found to be associated with large CNVs. Four genes affected by CNVs displayed a dosage effect in transcription and were probably affecting the growth and development of the potato plants [36]. FISH screening using subtracted random polymerase chain reaction (PCR) libraries as

Table 1: List of structural variations and their associations with phenotypes in plant species

Plant species	SVs identified		Genes covered	Accessions	Trait/genes associated with SVs	Method used for detection of SVs	Studies
	CNVs	PAVs					
Arabidopsis	1059 regions	–	500 genes	80 inbred lines	Adaptation to diverse environment	Re-sequencing and *de novo* assembly	Cao *et al.* [26]
	14 CNV events (comparing 16°C lineage with reference lineage at 22°C); 11 CNV events (comparing 28°C lineage with reference lineage at 22°C); 13 CNV events (comparing biotic stress lineage with reference lineage)	–	400 (comparing 16°C lineage with reference lineage at 22°C); 292 (comparing 28°C lineage with reference lineage at 22°C); 402 (comparing biotic stress lineage with reference lineage)	Three siblings from five lineage derived from common ancestor	Different environmental conditions (temperature and biotic stress)	CGH	DeBolt [16]
	2315 large InDels including CNVs		316 genes	Ler accession compared with Col0	Stress responsive genes	Re-sequencing	Lu *et al.* [27]
	1220 (Eil-0), 1312 (Lc-0), 1344 (Sav-0) and 987 (Tsu-1) genes with deletions were identified			Eil-0, Lc-0, Sav-0, Tsu-1 and Col-0 (reference)	Common ancestry and history of rearrangements	CGH and re-sequencing	Santuari *et al.* [28]
Barley	1 kb insertion in the upstream of the *HvAACT1* coding region			265 cultivated and 154 wild barley accessions were used to examine the presence of the specific insertion	Aluminium tolerance	Targeted re-sequencing	Fujii *et al.* [29]
	Four times *Bot1* copies in barley landrace Sahara 3771 as compared with clipper genotypes				Boron toxicity tolerance	Combination of mapping approaches, hybridization and qPCR	Sutton *et al.* [17]
Foxtail millet	37 232 SVs in SLX.Yugu1; 41 514 SVs SLX.Zhang gu		1612 SVs in genes in SLX.Yugu1; 2163 SVs in genes SLX.Zhang gu	Landrace, Shi-Li-Xiang (SLX) compared with the two reference genome sequences	–	Re-sequencing	Bai *et al.* [19]
Maize	>2000 regions	–	All CNVs present in genes	14 inbred lines	Disease response and heterosis	CGH	Beló *et al.* [20]
	10 000 segments	–	The majority (70%) of genes had an read-depth variants in at least one line	103 lines across pre-domestication and domesticated *Zea mays* lines	–	Re-sequencing	Chia *et al.* [30]
	–	296 genes putatively missing from one or more lines		Six elite maize inbred lines	–	Re-sequencing	Lai *et al.* [31]
	Tandem triplication of *MATE1* gene	–	*MATE1* gene	Three copy allele were identified from maize and teosinte diversity panel and validated in recombinant inbred lines	Aluminium tolerance	qPCR	Maron *et al.* [32]
	>400 segments	>1700	~50 genes associated with CNVs and 180 genes associated with PAVs	Mo17 and B73	Domestication	CGH	Springer *et al.* [7]
	3410 genes			19 inbred and 14 wild lines	–	CGH	Swanson-Wagner *et al.* [21]
	333 genes			278 inbred lines	Breeding selection	Re-sequencing	Jiao *et al.* [33]

(continued)

Table I: Continued

Plant species	SVs identified		Genes covered	Accessions	Trait/genes associated with SVs	Method used for detection of SVs	Studies
	CNVs	PAVs					
Opium	–	10 genes		Three varieties, F_2 population of 271 individuals	Noscapine synthesis	Re-sequencing	Winzer *et al.* [34]
Pigeonpea	–	29 regions	–	4 lines	Cytoplasmic male sterility	Re-sequencing	Tuteja *et al.* [35]
Potato	Four genes associated with CNVs			16 lines	Growth and development	FISH	Iovene *et al.* [36]
Rice	1676 segments	1327 genes	50% CNVs and all PAVs associated with genes	40 cultivated and 10 wild lines	Disease resistance and domestication	Re-sequencing	Xu *et al.* [22]
	641 segments	–	–	One line each from japonica and indica	–	CGH	Yu *et al.* [37]
Sorghum	17 111	16 487	CNVs associated with 2600 genes and PAVs associated with 1416 genes	Two sweet and one grain sorghum inbred lines	Disease resistance and selection	Re-sequencing	Zheng *et al.* [23]
Soybean	Significant levels of CNVs identified	25 genes		Williams 82 individuals and parental lines	Stress responsive genes	CGH	Haun *et al.* [38]
	–	18 600 regions	856 genes	14 cultivated and 17 wild lines	Metabolic and catalytic processes and disease resistance	Re-sequencing	Lam *et al.* [39]
	188–267 segments	133 regions	672 genes associated with CNVs	Archer, Minsor, Noir land Williams 82	Disease resistance and biotic stress	CGH	McHale *et al.* [24]
Wheat	Two to three copies of *Vrn1-A*	–	–	–	Flowering time	Targated re-sequencing	Díaz *et al.* [40]
	–	Deletion in upstream region of *Ppd-1* gene	–	–	Heading time	qPCR	Nishida *et al.* [25]
	85	7	–	–	Biotic and abiotic stresses	Re-sequencing	Saintenac *et al.* [41]

probes also provided the positions of microsatellite and chromosome-specific subtelomeric sequences [48]. Cytogenetically detectable heterochromatic variants have been used for species distinction and relationship studies in plants [49, 50]. These initial studies have provided knowledge of genome size variation that demonstrated the relatively consistent nature of genomes within a species. However, microscopic variations could be found even among closely related species, and these might be correlated with various adaptive features at the nuclear and organismic levels in plants. Microscopic variations in some genera occur in a discontinuous manner, forming groups of taxa, which are separated by regular time intervals. However, some genera showed continuous variation [49]. These facts demonstrated that microscopic genome variations could be used as corroborative evidence in plant systematics.

Submicroscopic SVs

Recent advances in DNA sequencing technology have allowed plant structural genetic variations to be analysed at a higher resolution than the microscopic studies described above. These SVs have been identified in either a genome-wide or a targeted manner, with varying degrees of resolution. Relatively little is known about genomic SVs and their association with phenotypic characteristics in plants. However, reports on such variants have started to appear (Table 1). Here we review recent SV studies in plant genomes.

Copy number variations

The term CNV is used to define sequences that demonstrate a variable copy number between individuals. The term has been used to describe duplications, deletions and insertions [51]. CNVs have been extensively characterized in maize [7]. In this study, genome-wide comparison of two inbred lines B73 and Mo17, identified 400 putative CNVs, and these CNVs were reported to be the result of tandem duplications [7]. In a subsequent study, genome-wide comparison of a set of 14 inbred maize lines identified thousands of CNVs [20]. In a further study in maize, CNVs were examined in 19 diverse inbred maize lines and 14 teosinte accessions [21]. This identified 479 genes with higher copy number and 3410 genes with fewer copies following comparison with a reference genome. Most of these CNVs were found to be present in related wild individuals, suggesting that these CNVs were not associated with deleterious genes responsible for lethality or major fitness loss [21].

In the small genome model plant *Arabidopsis*, CNVs were detected in 402 genes [16], while in rice, a comparison of japonica and indica cultivars identified 641 CNVs [37]. The majority of these rice CNVs suggested a loss of genomic segments in the indica cultivar 'Guang-lu-ai 4'. Japonica and indica rice diverged around 0.2–0.4 million years ago and display a high degree of DNA sequence variation [52]. Genome-wide patterns of CNVs have also been detected in sorghum by comparing two sweet and one grain inbred sorghum lines, identifying 3234 CNVs in 2600 genes [23]. Soybean was the first legume species to have its genome analysed for CNVs, and a total of 267 CNVs with an average size of 18–23 kb were detected across the genomes assayed [24] (Table 1).

The relationship between CNV occurrence and recombination frequency is not fully understood. In general, CNVs are scattered across plant genomes. Studies conducted in the maize genome have revealed that low-recombination regions such as telomeres show a greater number of CNVs [20, 21]. In contrast to maize, higher levels of CNV were identified in high-recombination regions in soybean and barley [18, 24].

Presence and absence variations

Sequences that are present in one genome and absent in another genome have been termed presence–absence variation (PAV). PAVs can be considered to be extreme CNVs, where the sequence is completely missing from one or more individual. A comparison of sequence data from two maize inbred lines (B73 and Mo17) detected 1783 PAVs that were present in the B73 genome and absent in the Mo17 genome. These PAVs relate to 1270 genes, suggesting that PAV affects a significant portion of maize genome. Analysis of these PAVs highlighted their association with ancestral evolution events and domestication [7]. Initially, CNVs and PAVs were combined for analysis of genome-wide variation in maize [21]. However, the mechanism of PAV formation was found to be different from that for CNVs and is not influenced by recombination. It was found that a short deletion mechanism that is based on short direct repeats likely contributes to the high rate of PAV among maize genotypes [53]. Comparing sequence data from sweet sorghum and grain sorghum lines identified 16 487 PAVs associated with 1416 genes. In pigeonpea (*Cajanus cajan*), PAVs have

been reported in the mitochondrial genomes of male-sterile (A-), maintainer (B-), hybrid (H-) and wild (W-) lines of pigeonpea [35]. Similar mitochondrial structural variations have been identified in other plant species including maize [54] and *Arabidopsis* [55].

Other structural variations

Other types of submicroscopic structural variation include inversions and translocations. These variations have been reported in nuclear and organelle genomes and are of considerable interest, as they can introduce novel diversity in plants. Several studies have reported the presence of subgenomic structural variations in mitochondrial genomes that have arisen from inversions and translocations [56, 57]. While such events in plant mitochondria increase organelle genome complexity, recombination has also been found to maintain genomic stability and may provide a mechanism to increase genetic variation in the absence of sexual reproduction [58]. Genomic inversions can be a driver of speciation, and this has been studied in plants using comparative genomics [59, 60]. An inverted region may not successfully recombine with its counterpart chromosome and might lead to infertility. Inversions are highly polymorphic in some species and may play a critical role in local adaptation [61]. Large-scale inversions have also been characterized in the chloroplast genomes of land plants [62]. Cytological studies have previously been conducted to characterize genomic inversions in various plant species; however, the application of large-scale genome sequencing will significantly help in characterizing the complex landscape of inversions and translocations in plant genomes.

APPROACHES TO IDENTIFY SUBMICROSCOPIC STRUCTURAL VARIATIONS

The on-going revolution in DNA sequencing technology known as NGS together with advances in bioinformatics have allowed structural genetic variations to be analysed at high resolution at a genome-wide level [63, 64]. SVs differ in size and complexity and hence different techniques have been used to characterize them in plant genomes. PCR-based approaches have been used for targeted regions of the genome. For example, real-time quantitative PCR (qPCR) was used to detect multiple copies of *Bot1* gene in barley genotypes [17], *MATE1* gene in maize genotypes [32] and a deletion in the upstream region of *Ppd-1* homeologs of wheat [25]. This technique offers a high sensitivity and a high-throughput alternative to the more traditional Southern blot used for determining gene copy number. PCR can also identify small translocations and inversions, as well as InDel polymorphism and CNVs [65]. Below we discuss approaches that have had a major impact on the discoveries of submicroscopic variants in the plant genome.

Microarrays

Microarray-based techniques were among the first used to detect genome-wide variation in human and plant genomes. Using array comparative genomic hybridization (aCGH), differentially labelled DNA from the test genome and a reference genome are hybridized to an array. Such an array contains thousands of probes developed from known gene sequences. BACs are the most popular arrayed targets in aCGH experiments, as they provide extensive coverage of the genome; however, cDNAs, PCR products and oligonucleotides can all be used as array targets. To increase the resolution of aCGH, the 'complexity' of the input DNA is reduced by a method called representation or whole-genome sampling [66]. A number of variations have been included in this approach to improve its efficiency, for instance using spotted oligonucleotides on Affymetrix arrays [67].

aCGH was first developed and applied for cancer genomics [14], and later used extensively in plant genomics to detect SVs [7, 16, 21, 24]. An early version of an array used in maize was composed of 14 423 BACs [7]. In comparison, the latest maize array contains 32 450 maize genes [21]. In *Arabidopsis*, a whole-genome CGH array was used to estimate SVs [16], and a recently developed high-resolution CGH platform was used to investigate the structure and diversity of genomic introgressions in two classical soybean near isogenic line populations [68]. Several factors affect aCGH-based SV detection. Gene distribution along the genome captured in arrays is not uniform, leading to bias; the majority of the probes are often designed to be complementary to a single genotype, reducing the efficiency of detecting SVs in other genotypes; sequences that are present in individuals and not in the reference sequence from which CGH arrays designed would not be represented; hybridization signals may deviate owing to DNA polymorphisms and lead to the false calling of SVs; and finally there remains a need to physically map the location of the probe in genome. A further challenge is applying moderate density arrays to highly repetitive plant

genomes. In this scenario, a high-density microarray platform designed for aCGH would greatly improve the efficiency of detection and estimation of SVs.

Evolving NGS techniques offer several advantages over aCGH by enabling the direct detection of DNA variations and recombination breakpoints [69]. NGS-based approaches also provide ability to detect inversions and translocations that are not generally detected by aCGH. However, aCGH would still be beneficial in genomic regions with multiple repeats where NGS-based assembly is difficult.

Genome sequencing/re-sequencing

In recent years, sequencing technologies have rapidly evolved from classical Sanger sequencing to NGS [70]. This has significantly lowered the cost of sequencing DNA. However, there are some limitations associated with these technologies such as the length of a DNA molecule that can be sequenced, though there are continuous improvements in this area. At present read lengths produced by the various technologies range from 25 bp to 15 kb. There is usually a compromise between read length, cost and accuracy, with low cost or longer read sequencing generally demonstrating significantly lower accuracy than some of the more popular technologies. The Illumina sequencing systems currently dominate the NGS market and they produce accurate reads of 150 bp for the HiSeq2500 and 300 bp for the MiSeq. Many NGS technologies such as those from Illumina use paired end or mate pair sequencing protocols, where two reads are generated with a known orientation and approximate distance between them. This significant assists the specificity of mapping or assembling this sequence data. Evolving technologies such as Single-Molecule Real Time (SMART) sequencing from Pacific Biosciences and Moleculo technology from Illumina have demonstrated the ability in reading long molecules of DNA up to 10 kb to 20 kb [71]. Nanopore technology also promises advances in this area, though little is known about the specific applications. Advances in DNA sequencing technology will continue to drive genomics and enhance the ability to detect structural variations with increasing resolution over a greater number of samples. There are three main approaches that can be used for the detection of SVs in plant genomes using DNA sequence data: (i) *de novo* assembly, (ii) re-sequencing approach and (iii) pan-genome.

i) The de novo assembly approach: In this approach two or more unique assemblies can be compared to identify and characterize SVs. Once the assemblies have been generated, this is a very efficient approach and can detect all types of SVs including CNVs, PAVs, translocations and inversions (Figure 1). The initial assembly needs high sequence coverage and sophisticated algorithms to reconstruct the genome from short overlapping sequences [72, 73]. This approach is the most robust for the characterization of SVs in a genome; however, the production of *de novo* assembled genomes of suitable quality remains the chief limitation. Draft plant genome assemblies are often highly fragmented and may contain many collapsed repeat regions that confound CNV detection. Improving and validating genome assemblies is an active research area, which is advancing through the application of novel algorithms and improved DNA sequence data. However, until the sequencing cost reduces significantly with substantially longer reads the *de novo* assembly of all genotypes representing a species is unfeasible and this approach is usually restricted to the detection of inter-species variation. Different draft genome assemblies from various plant species have been used to detect lineage and translocations and inversions [59, 60, 74].

ii) The re-sequencing approach: In the re-sequencing approach, DNA sequence reads from individual genotypes are aligned to a closely related reference genome (Figure 1). Differences between genomes then correlate to variations between the aligned reads and the reference genome. This approach can also be used for the detection of inversions, based on the orientation of aligned reads with the reference genome. Although this approach may not have such a high resolution as the *de novo* assembly approach, it will remain, in our opinion, the preferred method to detect intra-specific variation owing to its relatively low cost and lack of complexity associated with the generation of a *de novo* genome assembly for each variety. The re-sequencing approach has been used in sorghum, where a set of nearly 1500 genes differentiating sweet and grain sorghum were identified harbouring SVs [23]. Re-sequencing-based approaches are currently being applied to detect SVs in several other projects including the 1001 genome project in *Arabidopsis* [75], the maize panzea project (http://www.panzea.org) and the rice variation catalogue [22]. We are currently using this approach in pigeonpea, chickpea (*Cicer arietenum*) and peanut (*Arachis hypogaea*),

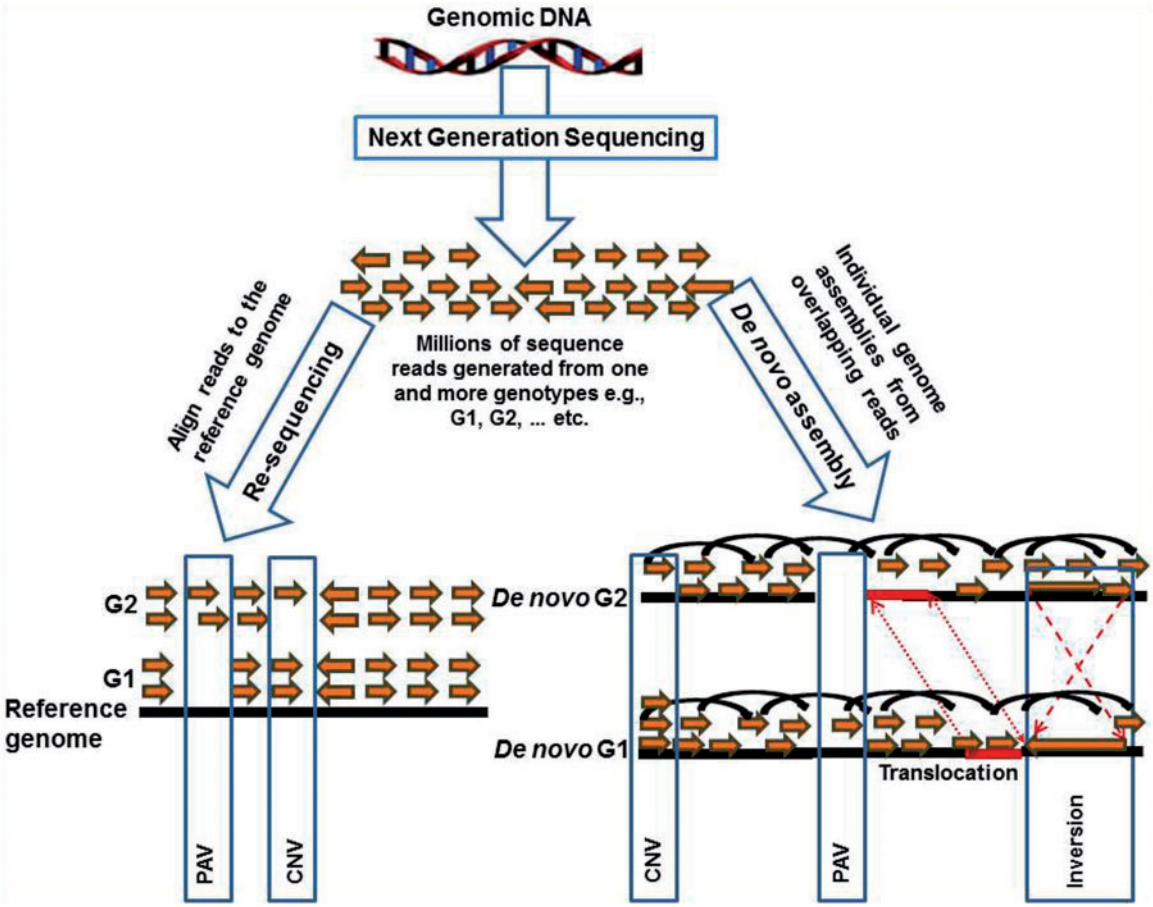

Figure 1: Two major NGS approaches to detect SVs are *de novo* assembly and re-sequencing. *De novo* assembly method is highly efficient to detect all types of SVs including CNVs, PAVs, inversions and translocations. Re-sequencing approaches are viable options to detect CNVs and PAVs.

re-sequencing 300 lines from reference sets for each species. These on-going efforts in a variety of plant species will provide insight into the distribution of SVs in plants as well as their evolution.

iii) The pan-genome: The pan-genome is composed of a core genome and a dispensable genome. The core genome contains genome segments or genes that are present in all accessions, while a dispensable genome is composed of partially shared and accession-specific DNA sequence elements. This concept of separate core and dispensable genomes was first described in prokaryotes [76]. A single genome sequence does not possess the entire genomic architecture of a species and so a pan-genome approach enables the description of a species rather than an individual at the genome level. Multiple accession sequencing projects in several plant species enables the creation of pan-genomes by defining the core and dispensable genome components of a species. The pan-genome has been described in some plants, e.g. maize [77–79] and *Arabidopsis thaliana* [80, 81].

ASSOCIATION OF SVs WITH PLANT PHENOTYPES

The role of SVs has been found to be important in human evolution and disease [13, 21], and SVs have been shown to be more frequent than SNPs in human genomes [13]. Although SVs have also been discovered in plants, their discovery and characterization are heavily reliant on the availability of at least one reference genome [82]. Few studies have been conducted to characterize the role of SVs in shaping plant phenotypes. The role of PAVs in determining plant phenotype has been demonstrated in opium (*Papaver somniferum*), where a cluster of 10 genes spanning a 221 kb genomic region were found to be associated with noscapine synthesis. Analysis of an F_2 mapping population indicated that these genes are tightly linked and absent in non-noscapine-producing lines [34]. Many of the CNVs identified in maize were found to be associated with domestication [21, 30]. The effect of selection on maize diversity has been estimated by sequencing 278

temperate maize inbred lines from different stages of breeding history. The results demonstrated that modern breeding has introduced highly dynamic genetic variations in the form of SNPs, InDels and CNVs, and affected a number of genic and non-genic regions in the maize genome [33]. The first-generation maize HapMap was constructed using sequence polymorphisms between 27 diverse inbred lines. This identified 18 regions that have undergone selective sweeps, including one region of 11 Mb on the long arm of chromosome 10 [83]. The second-generation maize HapMap was constructed using 103 lines and identified SVs that are enriched at loci associated with important traits [30]. An RNA-seq experiment using diverse lines of maize detected 757 loci that were restricted to a subset of the lines. Using *de novo* assembly of unmapped reads, novel transcripts were identified. It was also demonstrated that PAVs observed between different heterotic groups were transcribed. Furthermore, a core set and dispensable set of genes were identified [84]. Similarly Lai *et al*. [31] re-sequenced six elite maize inbred lines, including the parents of the commercial hybrids, and found 296 genes in B73 that were missing from at least one of the six inbred lines. Inbred lines representing different heterotic groups contained different sets of deleted genes. In both RNA-seq [84] and re-sequencing [31] studies it was postulated that unique transcripts or genes present in different heterotic groups might be contributing to the genetic basis of heterosis. In a recent study in maize by Maron *et al*. [32], CNVs were identified for the *MATE1* gene in aluminium-tolerant lines, but these were not common in teosinte. This study suggested that multiple copies of the *MATE1* gene arose recently and probably after domestication, and that CNVs were selected for their association with aluminium tolerance. *MATE1* expression found to be associated with CNV, where three *MATE1* copies were identical and part of a tandem triplication. Only three maize-inbred lines carrying the three-copy allele and demonstrating higher aluminium tolerance were identified from maize and teosinte diversity panels [32].

CNV of a 31 kb repeat segment observed in different haplotypes of the *Rhg1* locus encode multiple gene products in soybean cyst nematode (SCN)-resistant varieties. In SCN-susceptible varieties, one copy of the 31 kb segment per haploid genome was present. SCN resistance was found to be associated with increased expression of the CNV-related genes [85]. In an interesting study in palmer amaranth (*Amaranthus palmeri*), some plants were found resistant to herbicide glyphosate. These resistant plants contained 5–160 copies more of the *EPSPS* gene than susceptible plants. Expression and protein level of *EPSPS* gene was positively correlated with enhanced copy number [86].

In wheat, the recent association of SVs with plant phenotype has come in form of CNVs and large InDel polymorphisms. CNV for the gene *Vrn-A1* is associated with intermediate or late flowering phenotypes. CNV of *Ppd-B1* is found to contribute to photoperiod sensitivity in wheat [40]. Genotypes with a single copy of the *Ppd-B1* gene were photoperiod sensitive, while genotypes with elevated copy numbers were found to be early flowering and day-neutral [40]. An InDel polymorphism in the 50 bp upstream region of the *Ppd-1* gene was also associated with heading time of wheat cultivars [25]. In barley, a CACTA-like transposon insertion 5 kb upstream of the Open Reading Frame (ORF) of the aluminium tolerance gene *HcAACT1* enhances and alters the tissue localization of *HcAACT1* expression [29]. Another example of trait-associated CNVs in barley is the boron efflux carrier gene *Bot1* that plays an important role in boron tolerance [17]. CNVs have been found to be associated with nucleotide-binding leucine-rich repeat (NB-LRR) genes and receptor-like kinase (RLK) genes, known to be involved in plant defence-related mechanisms. CNVs related to disease resistance and biotic stress responses have also been identified in *Arabidopsis* [27], rice [22] and soybean [24]. Variable copies of these genes may be advantageous in the face of changing environmental conditions and possible threats posed by continuously evolving pest and pathogens.

OUTLOOK

Results from plant genome analysis have demonstrated the importance of SVs in evolutionary and biological processes. Initial studies conducted in a limited number of plant species suggest that a range of SVs are present and distributed across the genomes. It is anticipated that SVs will contribute an equal amount to the overall variation observed in the genome as SNPs. The low level of sequence diversity that is often suggested to exist in some of the self-pollinated or partially cross-pollinated crop

species might therefore be considered to be an overestimate. There remain challenges that need to be resolved before we achieve a complete understanding of the genome and its relationship with the plant phenotype. These include the effect of combinations of variants, interactions between genetic and environmental factors and epigenetic mechanisms. At present, no single method has the capability to detect the total complement of genomic structural variations. Even genome re-sequencing that is being applied in a number of important plant species would resolve only a proportion of the structural variation present in the genome. The highest resolution studies of SVs can be achieved by using a *de novo* assembly-based approach; however, this is not currently feasible for large numbers of individuals. Further, continuous improvements in sequencing technologies and reduction in costs will make it possible to detect nearly all variants between genomes. Even after *de novo* assembly, a significant amount of information could be lost owing to the challenges of assembling SVs using the available algorithms, and major advances in sequencing technology are required to facilitate accurate whole-genome assembly on a large scale. Improved assembly algorithms, combined with the ability to accurately sequence long stretches of DNA, would be beneficial to overcome many of these limitations. On-going and future efforts would greatly facilitate studies aimed at correlating genetic variations with plant performance. These efforts will also provide better understanding of the nature of the population history, natural selection and impact of structural variation in the plant genomes.

Key points

- This review describes recent reports of structural variations (SVs) in plant genomes and genomics technologies currently used to measure these SVs.
- Much of the recent attention in plant genetics is the result of the availability of high-resolution technologies for measuring these variants, including microarray-based techniques, and more recently, high-throughput DNA sequencing.
- On-going projects in a number of plant species promise to explore and characterize SVs and their associations with plant phenotypes.

FUNDING

Authors thank to United States Agency for International Development (USAID) and Department of Science and Technology (DST), Govt. of India under the framework of Australian Indo Strategic Research Funds (AISRF) for financial support. The Authors would like to acknowledge funding support from the Australian Research Council (Projects LP0882095, LP0883462 and LP110100200).

References

1. Edwards D, Batley J. Plant bioinformatics: from genome to phenome. *Trends Biotechnol* 2004;**22**:232–7.
2. Varshney RK, Ribaut JM, Buckler ES, *et al*. Can genomics boost productivity of orphan crops? *Nat Biotechnol* 2012;**30**:1172–6.
3. Imelfort M, Duran C, Batley J, *et al*. Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnol J* 2009;**7**:312–17.
4. Varshney RK, Nayak SN, May GD, *et al*. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol* 2009;**27**:522–30.
5. Berkman PJ, Lai K, Lorenc MT, *et al*. Next generation sequencing applications for wheat crop improvement. *Am J Bot* 2012;**99**:365–71.
6. Lai K, Duran C, Berkman PJ, *et al*. Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnol J* 2012;**10**:743–9.
7. Springer NM, Ying K, Fu Y, *et al*. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet* 2009;**5**:11.
8. Feuk L, Marshall CR, Wintle RF, *et al*. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet* 2006;**15**:R57–66.
9. Bignell GR, Santarius T, Pole JCM, *et al*. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res* 2007;**17**:1296–303.
10. Campbell PC, Stephens PJ, Pleasance ED, *et al*. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 2008;**40**:722–9.
11. Kolomietz E, Meyn MS, Pandita A, *et al*. The role of *Alu* repeat clusters as mediators of recurrent chromosomal aberrations in tumours. *Genes Chromosomes Cancer* 2002;**35**:97–112.
12. Kidd JM, Cooper GM, Donahue WF, *et al*. Mapping and sequencing of structural variation from eight human genomes. *Nature* 2008;**453**:56–64.
13. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med* 2010;**61**:437–55.
14. Raphael BJ. Structural variation and medical genomics. *PLoS Comput Biol* 2012;**8**:12.
15. Michael TP, Jackson S. The first 50 plant genomes. *Plant Genome* 2013;**6**:2.
16. DeBolt S. Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol Evol* 2010;**2**:441–53.
17. Sutton T, Baumann U, Hayes J, *et al*. Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science* 2007;**318**:1446–9.

18. Muñoz-Amatriaín M, Eichten SR, Wicker T, *et al*. Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol* 2013;**12**:14.
19. Bai H, Cao Y, Quan J, *et al*. Identifying the genome-wide sequence variations and developing new molecular markers for genetics research by re-sequencing a landrace cultivar of foxtail millet. *PloS One* 2013;**8**:9.
20. Belo A, Beatty MK, Hondred D, *et al*. Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor Appl Genet* 2010;**120**: 355–67.
21. Swanson-Wagner RA, Eichten SR, Kumari S, *et al*. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res* 2010;**20**:1689–99.
22. Xu X, Liu X, Ge S, *et al*. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* 2011;**30**:105–11.
23. Zheng LY, Guo XS, He B, *et al*. Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biol* 2011;**12**:R114.
24. McHale LK, Haun WJ, Xu WW, *et al*. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol* 2012;**159**:1295–308.
25. Nishida H, Yoshida T, Kawakami K, *et al*. Structural variation in the 5′ upstream region of photoperiod-insensitive alleles *Ppd-A1a* and *Ppd-B1a* identified in hexaploid wheat (*Triticum aestivum* L.), and their effect on heading time. *Mol Breed* 2013;**31**:27–37.
26. Cao J, Schneeberger K, Ossowski S, *et al*. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* 2011;**43**:956–63.
27. Lu P, Han X, Qi J, *et al*. Analysis of *Arabidopsis* genome-wide variations before and after meiosis and meiotic recombination by resequencing *Landsberg erecta* and all four products of a single meiosis. *Genome Res* 2012;**22**:508–18.
28. Santuari L, Pradervand S, Amiguet-Vercher AM, *et al*. Substantial deletion overlap among divergent *Arabidopsis* genomes revealed by intersection of short reads and tiling arrays. *Genome Biol* 2010;**12**:11.
29. Fujii M, Yokosho K, Yamaji N, *et al*. Acquisition of aluminium tolerance by modification of a single gene in barley. *Nat Commun* 2012;**3**:713.
30. Chia JM, Song C, Bradbury PJ, *et al*. Maize hapmap2 identifies extant variation from a genome in flux. *Nat Genet* 2012;**44**:803–807.
31. Lai J, Li R, Xu X, *et al*. Genome-wide pattern of genetic variation among elite maize inbred lines. *Nat Genet* 2010;**42**: 1027–30.
32. Maron LG, Guimarães CT, Kirst M, *et al*. Aluminum tolerance in maize is associated with higher *MATE1* gene copy number. *Proc Natl Acad Sci USA* 2013;**110**:5241–6.
33. Jiao Y, Zhao H, Ren L, *et al*. Genome-wide genetic change during modern breeding of maize. *Nat Genet* 2012;**44**:812–15.
34. Winzer T, Gazda V, He Z, *et al*. A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noncapine. *Science* 2012;**336**:1704–8.
35. Tuteja R, Saxena RK, Davila J, *et al*. Cytoplasmic male sterility-associated chimeric open reading frames identified by mitochondrial genome sequencing of four *Cajanus* genotypes. *DNA Res* 2013;**20**:485–495.
36. Iovene M, Zhang T, Lou Q, *et al*. Copy number variation in potato—an asexually propagated autotetraploid species. *Plant J* 2013;**75**:80–9.
37. Yu P, Wang C, Xu Q, *et al*. Detection of copy number variations in rice using array-based comparative genomic hybridization. *BMC Genomics* 2011;**12**:372.
38. Haun WJ, Hyten DL, Xu WW, *et al*. The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol* 2011;**155**: 645–55.
39. Lam HM, Xu X, Liu X, *et al*. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* 2010;**42**:1053–9.
40. Díaz A, Zikhali M, Turner AS, *et al*. Copy number variation affecting the photoperiod-*B1* and vernalization-*A1* genes is associated with altered flowering time in wheat (*Triticum aestivum*). *PLoS One* 2012;**7**:e33234.
41. Saintenac C, Jiang D, Akhunov ED. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol* 2011; **12**:R88.
42. Jacobs PA, Matsuura JS, Mayer M, *et al*. A cytogenetic survey of an institution for the mentally retarded: I. chromosome abnormalities. *Clin Genet* 1978;**13**:37–60.
43. Schwarzacher T, Leitch AR, Bennett MD, *et al*. In situ localization of parental genomes in a wide hybrid. *Ann Bot* 1989;**64**:315–24.
44. Chen CC, Chen CM, Hsu FC, *et al*. The pachytene chromosomes of maize as revealed by fluorescence in situ hybridization with repetitive DNA sequences. *Theor Appl Genet* 2000;**101**:30–6.
45. Lim KB, Wennekes J, de Jong JH, *et al*. Karyotype analysis of *Lilium longiflorum* and *Lilium rubellum* by chromosome banding and fluorescence in situ hybridization. *Genome* 2001;**44**:911–18.
46. Jiang J, Gill BS, Wang GL, *et al*. Metaphase and interphase fluorescence in situ hybridization mapping of the rice genome with bacterial artificial chromosomes. *Proc Natl Acad Sci USA* 1995;**92**:4487–91.
47. Kim JS, Childs KL, Islam-Faridi MN, *et al*. Integrated karyotyping of sorghum by in situ hybridization of landed BACs. *Genome* 2002;**45**:402–12.
48. Kato A, Lamb JC, Birchler JA. Chromosome painting using repetitive DNA sequence as probes for somatic chromosome identification in maize. *Proc Natl Acad Sci USA* 2004; **101**:13554–9.
49. Ohri D. Genome size variation and plant systematics. *Ann Bot* 1998;**82**:75–83.
50. Jong HD, Fransz JP, Zabel P. High resolution FISH in plants–techniques and applications. *Trends Plant Sci* 1999;**4**: 258–63.
51. Scherer SW, Lee C, Birney E, *et al*. Challenges and standards in integrating surveys of structural variation. *Nat Genet* 2007; **39**:7–15.
52. Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci USA* 2004;**101**: 12404–10.
53. Woodhouse MR, Schnable JC, Pedersen BS, *et al*. Following tetraploidy in maize, a short deletion mechanism

removed genes preferentially from one of the two homologs. *PLoS Biology* 2010;**8**:e1000409.

54. Allen JO, Fauron CM, Minx P, *et al*. Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. *Genetics* 2007;**177**:1173–92.
55. Davila JI, Arrieta-Montiel MP, Wamboldt Y, *et al*. Double-strand break repair processes drive evolution of the mitochondrial genome in *Arabidopsis*. *BMC Biol* 2011;**9**:64.
56. Nair CK. Mitochondrial genome organization and cytoplasmic male sterility in plants. *Journal of Biosciences* 1993;**18**: 407–22.
57. Mackenzie S, McIntosh L. Higher plant mitochondria. *Plant Cell* 1999;**11**:571–85.
58. Mach J. Cool as the cucumber mitochondrial genome: complete sequencing reveals dynamics of recombination, sequence transfer, and multichromosomal structure. *Plant Cell* 2011;**23**:2472.
59. Zhang G, Liu X, Quan Z, *et al*. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat Biotechnol* 2012;**30**: 549–54.
60. Varshney RK, Song C, Saxena RK, *et al*. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat Biotechnol* 2013;**31**:240–6.
61. Dobzhansky TG. *Genetics of the Evolutionary Process*. New York: Columbia University Press, 1970.
62. Kim KJ, Lee HL. Widespread occurrence of small inversions in the chloroplast genomes of land plants. *Mol Cells* 2005;**19**: 104–13.
63. Edwards D, Henry RJ, Edwards KJ. Preface: advances in DNA sequencing accelerating plant biotechnology. *Plant Biotechnol J* 2012;**10**:621–2.
64. Edwards D, Batley J, Snowdon R. Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet* 2013;**126**:1–11.
65. Wang D, Amornsiripanitch N, Dong X. A genomic approach to identify regulatory nodes in the transcriptional network of systemic acquired resistance in plants. *PLoS Pathogens* 2006;**2**:e123.
66. Kennedy GC, Matsuzaki H, Dong S, *et al*. Large-scale genotyping of complex DNA. *Nat Biotechnol* 2003;**21**: 1233–7.
67. Zhao X, Li C, Paez JG, *et al*. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res* 2004;**64**: 3060–71.
68. Stec AO, Bhashkar PB, Bolon YT, *et al*. Genomic heterogeneity and structural variation in soyabean near isogenic lines. *Front Plant Sci* 2013;**4**:104.
69. Chen W, Kalscheuer V, Tzschach A, *et al*. Mapping translocation breakpoints by next-generation sequencing. *Genome Res* 2008;**18**:1143–9.
70. Thudi M, Li Y, Jackson SA, *et al*. Current state-of-art of sequencing technologies for plant genomics research. *Brief Funct Genomics* 2012;**11**:3–11.
71. Feuillet C, Leach JE, Rogers J, *et al*. Crop genome sequencing: lessons and rationales. *Trends Plant Sci* 2011;**16**: 77–88.
72. Imelfort M, Batley J, Grimmond S, *et al*. Genome sequencing approaches and successes. In: Somers D, Langridge P, Gustafson J (eds). *Plant Genomics*. USA: Humana Press, 2009:345–58.
73. Imelfort M, Edwards D. *De novo* sequencing of plant genomes using second-generation technologies. *Brief Bioinform* 2009;**10**:609–18.
74. Varshney RK, Chen W, Li Y, *et al*. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat Biotechnol* 2012;**30**:83–9.
75. Weigel D, Mott R. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol* 2009;**10**:107.
76. Tettelin H, Masignani V, Cieslewicz MJ, *et al*. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial ''pan-genome''. *Proc Natl Acad Sci USA* 2005;**102**:13950–5.
77. Morgante M, Brunner S, Pea G, *et al*. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* 2005;**37**:997–1002.
78. Brunner S, Pea G, Rafalski A. Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize. *Plant J* 2005;**43**:799–810.
79. Hirsch CN, Foerster JM, Johnson JM, *et al*. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* 2014; **26**:121–35.
80. Polak P, Domany E. *Alu* elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* 2006;**7**:133.
81. Johnson R, Gamblin RJ, Ooi L, *et al*. Identification of the REST regulon reveals extensive transposable element mediated binding site duplication. *Nucleic Acids Res* 2006; **34**:3862–77.
82. Morgante M, De Paoli E, Radovic S. Transposable elements and the plant pangenomes. *Curr Opin Plant Biol* 2007; **10**:149–55.
83. Gore MA, Chia JM, Elshire RJ, *et al*. A first-generation haplotype map of maize. *Science* 2009;**326**:1115–17.
84. Hansey CN, Vaillancourt B, Sekhon RS, *et al*. Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. *PLoS One* 2012;**7**:e33071.
85. Cook DE, Lee TG, Guo X, *et al*. Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science* 2012;**338**:1206–9.
86. Gaines TA, Zhang W, Wang D, *et al*. Gene amplification confers glyphosate resistance in Amaranthus palmeri. *Proc Natl Acad Sci USA* 2010;**107**:1029–34.